On Perceptual Distortion Minimization and Nonlinear Least-Squares Frequency Estimation

Christensen, Mads Græsbøll; Jensen, Søren Holdt

Published in:
IEEE transactions on speech and audio processing

DOI (link to publication from Publisher):
10.1109/TSA.2005.860347

Publication date:
2006

Document Version
Accepted author manuscript, peer reviewed version

On Perceptual Distortion Minimization and Nonlinear Least-Squares Frequency Estimation

Mads Græsbøll Christensen*, *Student Member, IEEE*, and Søren Holdt Jensen, *Senior Member, IEEE*

***Abstract*— In this paper, we present a framework for perceptual error minimization and sinusoidal frequency estimation based on a new perceptual distortion measure and we state its optimal solution. Using this framework, we relate a number of well-known practical methods for perceptual sinusoidal parameter estimation such as the pre-filtering method, the weighted matching pursuit and the perceptual matching pursuit. In particular, we derive and compare the sinusoidal estimation criteria used in these methods. We show that for the sinusoidal estimation problem, the pre-filtering method and the weighted matching pursuit are equivalent to the perceptual matching pursuit under certain conditions.**

I. Introduction

THE problem of estimating the parameters of a set of sinusoids in noise arises in many different applications. In digital processing of speech, the sinusoidal estimation problem arises in such applications as speech modeling and coding [1]–[5] and speech enhancement [6] and more recently, renewed interest in sinusoidal coding of speech has been spurred by the increasing interest in voice over packet-based networks [7]–[10]. Also in the field of audio processing, the sinusoidal signal model has been of interest for music analysis and synthesis [11]–[13], and parametric coding of audio [14]–[20]. In speech and audio processing the sinusoids can be seen as a parametric representation of the quasi-periodic, i.e. tonal, signal components, while the noise can be seen as the unvoiced, stochastic signal components [13]. The latter could, for example, be unvoiced speech, the bow noise of a violin, quantization errors or processing noise.

The applications mentioned above have in common that it is of interest to find a compact representation, or in other words to represent the signal in as few, physically meaningful parameters as possible. Since the end receiver of these signals is the human auditory system, it is also of interest to represent the perceptually most important components. In audio coding in particular, it is of interest to estimate and transmit only the parameters of audible sinusoids and in recent years, much effort has been put into this problem. Many different methods for solving this have been proposed, e.g. [21]–[28] all implement this in what seem to be different ways. Often, these methods rely heuristic rules taken from psychoacoustic experiments, while estimation theory, on the other hand, relies on statistical signal processing in finding model parameters. In [25] sinusoidal components are found in an iterative manner by assigning a perceptual weight to the spectrum and then picking the most dominant peak of the weighted spectrum. Another method is the so-called pre-filtering method, where the observed signal is filtered using a perceptual filter in order to achieve a weighting of the sinusoidal components, c.f. [26]. The methods of [27] and [28] are different methods yet—they rely on loudness and excitation pattern similarity criteria for sinusoidal component selection, respectively.

In coding applications it is of particular interest to state the estimation criterion in a way that defines a distortion measure or metric. A globally optimal solution that minimizes this distortion measure ensures that at a given bit-rate (for a certain number of sinusoids in the case of sinusoidal coding), the lowest possible distortion is achieved. When the distortion measure is a perceptual one, meaning that it reflects the human auditory system, we can then claim that the perceived distortion is minimized at the given bit-rate. In linear predictive speech coding, for example, perception is traditionally taken into account using a fairly simple approach, where the noise spectrum is shaped by a perceptual weighting filter, which is derived directly from the linear prediction filter of the speech signal [29].

A recently published psychoacoustic masking model for audio coding has been shown to form a distortion measure [30], [31], and this distortion measure has been applied successfully to the sinusoidal estimation problem in [15], [23], [32], [33]. Based on this we define the perceptual frequency estimation problem and its optimal solution. We then analyze and relate a number of different practical perceptual frequency estimators that are all based on least-squares in this framework. In particular, we study the estimation criteria of these estimators. This allows us to analyze, quantify and understand the nature of the approximations made in these estimators. An important result is that the estimation criteria of the pre-filtering method and the weighted matching pursuit can be derived from the perceptual matching pursuit from the same assumption. Since many applications rely on a physical interpretation of the estimated parameters, the statistical properties of the estimators in question are also of significant importance. In that spirit we also investigate how the least-squares based estimators relate to estimation theory and maximum likelihood frequency estimation.

The rest of this paper is organized as follows. In Section II the frequency estimation problem is introduced along with the nonlinear least-squares frequency estimator. Then, in Section III, we relate this to a simpler, common estimator, namely matching pursuit. In Section IV we proceed to introduce a perceptual distortion measure that can be written in the form

The authors are with the Department of Communication Technology, Aalborg University, Denmark (email: {mgc,shj}@kom.aau.dk, homepage: http://kom.aau.dk/~{mgc,shj}).

This research was supported by the ARDOR (Adaptive Rate-Distortion Optimized sound codeR) project, EU grant no. IST–2001–34095, and the Intelligent Sound project, Danish Technical Research Council grant no. 26–02–0092.

of a circulant, symmetric perceptual weighting matrix. In Section V we use this measure to formulate the perceptual frequency estimation problem and its optimal solution in terms of the perceptual nonlinear least-squares estimator. Moreover, we relate this to an approximation, namely the perceptual matching pursuit. The eigenvalue decomposition (EVD) of the perceptual weighting matrix and approximations with application to the problem at hand are studied in Section VI. In Section VII we then show how this can be used to relate a number of well-known perceptual sinusoidal frequency estimators. We present some illustrative numerical examples in Section VIII, and we summarize the results and give conclusions in Sections IX and X, respectively.

II. The Frequency Estimation Problem

The basic problem addressed in this paper can be stated as follows. Given a real observed signal $x(n)$ for $n = 0, \ldots, N-1$, find the parameters of the signal of interest $\hat{x}(n)$ in additive noise $e(n)$:

$$x(n) = \hat{x}(n) + e(n). \tag{1}$$

In our case the signal of interest $\hat{x}(n)$ is a sum of sinusoidal components

$$\hat{x}(n) = \sum_{l=1}^{L} A_l \cos\left(\omega_l n + \phi_l\right), \tag{2}$$

with each component having a constant amplitude A_l, initial phase ϕ_l, and frequency ω_l. The problem is then to estimate these parameters, in particular the frequencies $\boldsymbol{\omega} = [\ \omega_1\ \cdots\ \omega_L\ ]^T$. In the same process, the amplitudes and phases are usually also found, but as we shall see, these can written as complex linear parameters and can then be found in straightforward way.

Supposing that $e(n)$ is zero-mean white, i.i.d. (independent and identically distributed over observations) Gaussian noise of variance σ^2, the likelihood function $p(\mathbf{x};\boldsymbol{\omega})$, which is a function of the observed signal and the model parameters (here only the frequencies) can be written as (see e.g. [34])

$$\begin{aligned} p(\mathbf{x};\boldsymbol{\omega}) &= \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left|x(n)-\hat{x}(n)\right|^2\right] \\ &= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\left|x(n)-\hat{x}(n)\right|^2\right]. \end{aligned} \tag{3}$$

Introducing a vector containing the observed signal $\mathbf{x} = [\ x(0)\ \cdots\ x(N-1)\ ]^T$ and a vector containing the modeled signal $\hat{\mathbf{x}} = [\ \hat{x}(0)\ \cdots\ \hat{x}(N-1)\ ]^T$, this can be written as

$$p(\mathbf{x};\boldsymbol{\omega}) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\|\mathbf{x}-\hat{\mathbf{x}}\|_2^2\right]. \tag{4}$$

Taking the logarithm, we get the log-likelihood function

$$\ln p(\mathbf{x};\boldsymbol{\omega}) = -\frac{N}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\|\mathbf{x}-\hat{\mathbf{x}}\|_2^2. \tag{5}$$

We see that for white Gaussian noise, maximizing the likelihood function is the same as minimizing the squared error between the observed signal and the signal model. In the nonlinear least-squares frequency estimator (NLS), the sinusoidal frequencies are estimated by minimizing exactly this error in a least-squares sense. The method is known as nonlinear least-squares as the cost function is nonlinear in the unknown frequencies. It is interesting, but perhaps not surprising, that in this particular case, the statistical approach of maximum likelihood (ML) turns into a deterministic method that matches the signal model to the outcome of the random process. The resulting estimator can be stated as the solution to the following problem [35]:

$$\min \|\mathbf{x}-\hat{\mathbf{x}}\|_2^2 = \min \|\mathbf{x}-\mathbf{Z}\mathbf{a}\|_2^2. \tag{6}$$

Here, the matrix $\mathbf{Z} \in \mathbb{C}^{N\times 2L}$ ($N > 2L$) is a so-called Vandermonde matrix[1] defined as

$$\mathbf{Z} = \begin{bmatrix} z_1^0 & z_1^{-0} & \cdots & z_L^0 & z_L^{-0} \\ z_1^1 & z_1^{-1} & \cdots & z_L^1 & z_L^{-1} \\ \vdots & \vdots & & \vdots & \vdots \\ z_1^{N-1} & z_1^{-(N-1)} & \cdots & z_L^{N-1} & z_L^{-(N-1)} \end{bmatrix}, \tag{7}$$

where signal poles $z_l = \exp(j\omega_l)$ come in complex conjugate pairs. Assuming that the signal poles are distinct, the matrix has full rank. Furthermore, we have that $\mathbf{a} \in \mathbb{C}^{2L}$, $\mathbf{a} = [\ a_1\ a_1^*\ \cdots\ a_L\ a_L^*\ ]^T$ with

$$a_l = \frac{A_l}{2}\exp(j\phi_l). \tag{8}$$

The NLS frequency estimates are then the combination of L frequencies (with $\hat{}$ denoting estimates) that minimizes the squared error, i.e.,

$$\hat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega}}{\operatorname{argmin}} \|(\mathbf{x}-\mathbf{Z}\mathbf{a})\|_2^2. \tag{9}$$

This can be formulated as a maximization problem using the principle of orthogonality:

$$\begin{aligned} \hat{\boldsymbol{\omega}} &= \underset{\boldsymbol{\omega}}{\operatorname{argmin}}\, \mathbf{x}^H\mathbf{x} - \mathbf{x}^H\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{x} && (10) \\ &= \underset{\boldsymbol{\omega}}{\operatorname{argmax}}\, \mathbf{x}^H\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{x}. && (11) \end{aligned}$$

The corresponding amplitude and phase estimates are the solution to (6) given the frequencies:

$$\hat{\mathbf{a}} = \left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{x}. \tag{12}$$

For more on estimation of amplitudes and phases, we refer the reader to the study in [37]. In order to solve the frequency estimation problem this way, we have to search (numerically) for the combination of the L complex sinusoids that minimize the 2-norm of the error signal. This is essentially the subspace pursuit of [38] with the sum of sinusoids being the target subspace. Clearly, this is a complex procedure and it is not easily solved. In most real-time applications, solving this problem directly is not feasible. For more on the intractability of this problem, we refer the reader to [39].

One may argue that this point of view is unrealistic both in terms of solving the problem optimally and in terms of the assumptions with respect to the noise, but the NLS frequency estimator is very interesting from a theoretical point of view

[1] Vandermonde matrices are sometimes defined to be square [36].

because it has excellent statistical performance. For the white Gaussian noise case, it is efficient and unbiased—it attains the Cramér-Rao Bound (see e.g. [35], [40], [41]).

In speech and audio processing the noise cannot generally be assumed to be white. For the colored noise case, with the Gaussian noise $e(n)$ now having the positive definite (non-diagonal) covariance matrix $\mathbf{\Sigma}$, the likelihood function is [41]

$$p(\mathbf{x};\boldsymbol{\omega}) = Q \exp\left[-\frac{1}{2}\left(\mathbf{x}-\hat{\mathbf{x}}\right)^H \mathbf{\Sigma}^{-1}\left(\mathbf{x}-\hat{\mathbf{x}}\right)\right], \quad (13)$$

with

$$Q = \frac{1}{(2\pi)^{\frac{N}{2}}\sqrt{\det(\mathbf{\Sigma})}}. \quad (14)$$

The corresponding maximum likelihood estimator is then

$$\hat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega}}{\operatorname{argmin}} \left(\mathbf{x}-\hat{\mathbf{x}}\right)^H \mathbf{\Sigma}^{-1}\left(\mathbf{x}-\hat{\mathbf{x}}\right). \quad (15)$$

Without prior knowledge of the noise covariance matrix $\mathbf{\Sigma}$, this problem is clearly more difficult to solve than for white noise where $\mathbf{\Sigma} = \sigma^2\mathbf{I}$ and $\det(\mathbf{\Sigma}) = \sigma^{2N}$. However, as shown in [41], the NLS estimator in (11) is also asymptotically efficient for colored noise under some mild conditions. For more details on the relation between the NLS and ML estimators for the colored noise case and the associated Cramér-Rao bound, we refer the reader to [41], and for a practical method that achieves the Cramér-Rao bound see [42]. For non-Gaussian noise, the NLS estimator loses its maximum likelihood interpretation [41]. Here it must be stressed that we are not arguing as to the nature of noise in audio signals but rather as to the optimality of some commonly used methods that are based on least-squares.

III. Relaxation of the NLS Estimator

In this section we treat the relationship between the NLS frequency estimator and a well-known method for sinusoidal parameter estimation, namely matching pursuit [43]. As we shall see, there is a close relation between the two, although originally proposed in two entirely different contexts.

In matching pursuit a signal model is built iteratively by solving for one component at a time. This is done by finding the component from a dictionary, in this case composed of a set of complex sinusoids of different frequencies, that minimizes some norm (here the 2-norm) of the residual, which is formed by subtracting the i-th component from the i-th residual, i.e.,

$$r_{i+1}(n) = r_i(n) - \hat{A}_i\cos(\hat{\omega}_i n + \hat{\phi}_i), \quad (16)$$

with the residual being initialized as $r_1(n) = x(n)$. The Vandermonde matrix $\mathbf{Z}$ now contains the vector $\mathbf{z} = \left[\begin{array}{cccc} \exp(j\omega 0) & \cdots & \exp(j\omega(N-1)) \end{array}\right]^T$ and its complex-conjugate:

$$\mathbf{Z} = \left[\begin{array}{cc} \mathbf{z} & \mathbf{z}^* \end{array}\right]. \quad (17)$$

The frequency is then estimated as the minimizer of the 2-norm of the residual at iteration $i+1$

$$\hat{\omega}_i = \underset{\omega}{\operatorname{argmin}} \|\mathbf{r}_{i+1}\|_2^2 = \underset{\omega}{\operatorname{argmin}} \|\mathbf{r}_i - \mathbf{Z}\mathbf{a}\|_2^2 \quad (18)$$

$$= \underset{\omega}{\operatorname{argmax}} \mathbf{r}_i^H \mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{r}_i, \quad (19)$$

where $\mathbf{r}_i = \left[\begin{array}{ccc} r_i(0) & \cdots & r_i(N-1) \end{array}\right]^T$. After i iterations, the signal model is simply

$$\hat{x}_i(n) = \sum_{l=1}^{i} \hat{A}_l\cos(\hat{\omega}_l n + \hat{\phi}_l). \quad (20)$$

Writing out the estimation criterion (19) (here denoted J), we get

$$J = \mathbf{r}_i^H \mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{r}_i \quad (21)$$

$$= \mathbf{r}_i^H \left[\begin{array}{cc} \mathbf{z} & \mathbf{z}^* \end{array}\right] \left[\begin{array}{cc} \mathbf{z}^H\mathbf{z} & \mathbf{z}^H\mathbf{z}^* \\ \mathbf{z}^T\mathbf{z} & \mathbf{z}^H\mathbf{z} \end{array}\right]^{-1} \left[\begin{array}{c} \mathbf{z}^H \\ \mathbf{z}^T \end{array}\right] \mathbf{r}_i. \quad (22)$$

We see that this is still a subspace pursuit, but in this case the subspace is a function of one variable ω. This is sometimes referred to as a conjugate-subspace pursuit [38]. Assuming that the complex sinusoid and its complex-conjugate are well separated in frequency (not close to 0 or π relative to N), the inner product between the two can be assumed to be zero[2]:

$$\mathbf{z}^H\mathbf{z}^* \approx 0. \quad (23)$$

The estimation criterion (22) can then be reduced significantly:

$$J = \mathbf{r}_i^H \left[\begin{array}{cc} \mathbf{z} & \mathbf{z}^* \end{array}\right] \left[\begin{array}{cc} \mathbf{z}^H\mathbf{z} & 0 \\ 0 & \mathbf{z}^H\mathbf{z} \end{array}\right]^{-1} \left[\begin{array}{c} \mathbf{z}^H \\ \mathbf{z}^T \end{array}\right] \mathbf{r}_i \quad (24)$$

$$= 2\frac{\left|\mathbf{z}^H\mathbf{r}_i\right|^2}{\mathbf{z}^H\mathbf{z}}. \quad (25)$$

The sinusoidal frequency estimation criterion can now be written in the well-known form

$$\hat{\omega}_i = \underset{\omega}{\operatorname{argmax}} \frac{|\langle \mathbf{z}, \mathbf{r}_i\rangle|^2}{N}, \quad (26)$$

with $< \cdot, \cdot >$ denoting the inner product. The associated optimum complex scaling is

$$\hat{a}_i = \frac{\langle \mathbf{z}, \mathbf{r}_i\rangle}{N}, \quad (27)$$

which relates to the amplitude and phase in (16) as described in (8). We see that for the case of a sinusoidal dictionary MP is the NLS estimator in the one sinusoid case. It can be solved efficiently since the inner products $\langle \mathbf{z}, \mathbf{r}_i\rangle$ can be found using FFTs. Clearly, matching pursuit is a simplified approximation to (11). It can be seen as a relaxation of the original problem, where instead of solving the multidimensional nonlinear problem, we break it into several one-dimensional minimizations that have efficient implementations. Matching pursuit converges in the respective norm as i grows and the distortion is a non-increasing function of i (see [43]). It does not, generally, converge to zero in a finite number of iterations for the sinusoidal case as later iterations may introduce new spectral components due to the non-orthogonality of the components of redundant dictionaries. Sometimes this is also referred to as the readmission problem [44]. There are several ways to compensate for these problems (see for example [39], [44]–[47]).

[2]For the 2-norm case considered here, the conjugate-subspace pursuit can be solved efficiently without this assumption. However, this is not the case for the methods considered later in this paper.

On a historical note, the estimation procedure of [5], [11] first introduced in [48] is similar to that of matching pursuit for complex sinusoids later introduced in [43]. The RELAX algorithm [42] is an iterative sinusoidal frequency estimation algorithm, where the efficient solution to the one-sinusoid estimation problem is exploited in a recursive manner. It has been demonstrated to have excellent statistical performance achieving the Cramér-Rao bound for both white and colored Gaussian noise [41].

IV. A Perceptual Distortion Measure

It is well-known that the 2-norm error measure does not correlate well with human sound perception. The choice of a distortion measure involves a trade-off between many factors. On one hand we would like to have a measure that takes as much of the processing in the human auditory into account as possible, while on the other hand we would like to have a measure which defines a mathematical norm. Another desirable property of the measure is that it can be incorporated in an efficient algorithm. A generalized perceptually weighted 2-norm can be written as

$$\left\|\mathbf{W}\left(\mathbf{x}-\hat{\mathbf{x}}\right)\right\|_2^2, \tag{28}$$

where $\mathbf{W}$ is a so-called perceptual weighting or sensitivity matrix (e.g. [25], [49]). Even very sophisticated distortion measures can be expressed this way. For example, in [49] the model of [50], [51] is linearized and put into the form of (28). Since we are here concerned with the estimation of stationary sinusoids, we assume the observed signal to be stationary. For stationary signals, the masking in the human auditory system is predominantly caused by simultaneous masking. Masking analysis in audio coding usually only considers distortions in the individual auditory filters, see e.g. ISO 11172-3 (MPEG-1) Psychoacoustic Model 1 described in [52]. Recently, it has been shown that significant improvements are gained by taking spectral integration into account [30], [31]. Using the masking model proposed in [30], [31], which was derived specifically for sinusoidal coding, the distortion D for a particular segment can be written as

$$D=\int_{-\pi}^{\pi} A(\omega)|E(\omega)|^2 d\omega, \tag{29}$$

where $A(\omega)$ is a real, positive perceptual weighting function and $E(\omega)$ is the discrete-time Fourier transform of the error $e(n)=w(n)\left[x(n)-\hat{x}(n)\right]$ where $w(n)$ is the analysis window. When the weighting function is chosen as the reciprocal of the masking threshold, the error spectrum which results from minimizing D will be shaped like the masking threshold.

In the coming analyses, we assume a rectangular window ($w(n)=1\ \forall n$) for simplicity and mathematical convenience since we shall rely on asymptotic properties. In practice, the weighting function $A(\omega)$ and the error spectrum $E(\omega)$ are uniformly sampled spectra $A(k)$ and $E(k)$, respectively, and the integral (29) can be calculated as a summation of point-wise multiplications in the frequency domain:

$$D=\sum_{k=0}^{K-1}|\sqrt{A(k)}E(k)|^2. \tag{30}$$

The point-wise spectral multiplication corresponds to circular convolution in the time-domain, i.e.

$$\sum_{m=0}^{K-1} h(m)e((k-m)\,(\mathrm{mod}\,K))\leftrightarrow\sqrt{A(k)}E(k), \tag{31}$$

with $\leftrightarrow$ denoting Fourier transform pairs. Furthermore, from Parseval's theorem, we have that the inner product can be calculated in the frequency domain as

$$\sum_{n=0}^{K-1} x^*(n)y(n)=\frac{1}{K}\sum_{m=0}^{K-1} X^*(k)Y(k). \tag{32}$$

This means that the discrete distortion measure (30) can be written as the 2-norm of a circular convolution:

$$D=\sum_{k=0}^{K-1}|\sum_{m=0}^{K-1} h(m)e((k-m)\,(\mathrm{mod}\,K))|^2. \tag{33}$$

The sampling frequency of the reciprocal of the masking curve $A(k)$ (and thus the length of the corresponding filter) is determined by the human auditory system and not by the input signal.

The distortion measure can now be put into the more convenient matrix-vector notation:

$$D=\|\mathbf{He}\|_2^2 \tag{34}$$

with $\mathbf{H}$ being the perceptual weighting matrix, in this case a filtering matrix, having the following structure

$$\mathbf{H}=\begin{bmatrix} h(0) & h(K-1) & \cdots & h(1) \\ h(1) & h(0) & \cdots & h(K-1) \\ \vdots & \vdots & \ddots & \vdots \\ h(K-1) & h(K-2) & \cdots & h(0) \end{bmatrix}, \tag{35}$$

and $\mathbf{e}=\left[\ e(0)\ \cdots\ e(K-1)\ \right]^T$. This means that there is a duality between the spectral distortion measure and the two-norm of the circularly filtered error signal. This interpretation offers insights into the relation between a number of methods for perceptual frequency estimation. We will return to this later in the paper.

We now discuss how to derive an appropriate filter from the perceptual weighting function $A(k)$. As the perceptual filter has to be derived for each segment, computational complexity is of considerable importance. The simplest solution is to compute the impulse response as the inverse Fourier transform of $\sqrt{A(k)}$ for $n=0,\ldots,K-1$, i.e.,

$$h(n)=\frac{1}{K}\sum_{k=0}^{K-1}\sqrt{A(k)}\exp\left(j2\pi kn/K\right) \tag{36}$$

$$=\frac{1}{K}\sum_{k=0}^{K-1}\sqrt{A(k)}\cos\left(2\pi kn/K\right), \tag{37}$$

where the last line follows from $A(k)$ being real and symmetric ($A(k)=A(K-k)$), which also means that $h(n)$ is symmetric, i.e. $h(n)=h(K-n)$. This procedure leaves us with an impulse response of length K while our observed signal is of length N. Typically, the required length of the spectral weighting function is higher than the number of time-samples, i.e. $N<K$. The signal and model vectors can

then easily be zero-padded to length K or the last $K-N$ columns of $\mathbf{H}$ can be truncated. Filters of arbitrary order can be obtained using standard methods, and in the following sections we assume that the impulse response has been derived such that it has length N.

V. Perceptual NLS and MP

In many applications such as audio modeling and coding, it is of interest to extract only the perceptually most relevant sinusoidal component of the observed signal. Indeed, in audio coding, where the problem can be stated as minimizing the perceived distortion given some rate constraint, convergence in the perceptual distortion as we increase the number of sinusoids (and thus the rate) is desirable. Using the definitions in Section IV, we can restate the NLS frequency estimator as the following perceptually meaningful least-squares problem

$$\min \|\mathbf{H}(\mathbf{x}-\hat{\mathbf{x}})\|_2^2. \tag{38}$$

Let $\boldsymbol{\omega} = [\ \omega_1\ \cdots\ \omega_L\ ]^T$ be the set of frequencies that describe the Vandermonde matrix $\mathbf{Z} \in \mathbb{C}^{N\times 2L}$. Then the perceptual NLS estimates of the frequencies (and the corresponding optimal amplitudes and phases) are the solution to the problem

$$\hat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega}}{\operatorname{argmin}} \|\mathbf{H}(\mathbf{x}-\mathbf{Z}\mathbf{a})\|_2^2. \tag{39}$$

The vector $\hat{\boldsymbol{\omega}}$ is the vector containing the set of the frequencies of L sinusoids that minimize the filtered, weighted 2-norm and the vector $\hat{\mathbf{a}}$ contains the amplitudes and phases of those sinusoids in polar form. Since the filtering matrix is real and symmetric, i.e. $\mathbf{H}^H\mathbf{H} = \mathbf{H}^2$, these can be estimated as

$$\hat{\mathbf{a}} = \left(\mathbf{Z}^H\mathbf{H}^2\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{H}^2\mathbf{x}. \tag{40}$$

Substituting this into (39), we get

$$\hat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega}}{\operatorname{argmin}} \|\mathbf{H}(\mathbf{x}-\mathbf{Z}\mathbf{a})\|_2^2 \tag{41}$$

$$= \underset{\boldsymbol{\omega}}{\operatorname{argmax}}\, \mathbf{x}^H\mathbf{H}^2\mathbf{Z}\left(\mathbf{Z}^H\mathbf{H}^2\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{H}^2\mathbf{x}. \tag{42}$$

This re-statement of the NLS frequency estimator allows us to estimate only the perceptually significant sinusoids and disregard inaudible ones, and to find the amplitudes and phases in such a way that artifacts are not introduced in the decoded signal. This formulation is only relevant when we are interested in a subset of the total number of sinusoids. Otherwise, there is no need for the spectral weighting of the error in the frequency estimation. However, the total number of sinusoids is generally unknown and robustness with respect to the number of sinusoids is desirable. We mention in passing that it also may be advantageous to incorporate the perceptual distortion in the estimation of amplitudes and phases as in (40) since erroneous estimates may introduce components in parts of the spectrum where no masker is present.

In terms of projections and transformations, the filtering matrix $\mathbf{H}$ can be thought of as a transformation to a perceptual domain and the problem of finding the optimal signal model can be seen as a projection problem. Then, the transformed input signal is projected orthogonally onto the column space of the transformed signal model. This introduces an error which is orthogonal to the signal model in the perceptual domain. This is illustrated in Figure 1 with $\mathcal{R}(\cdot)$ denoting the range.

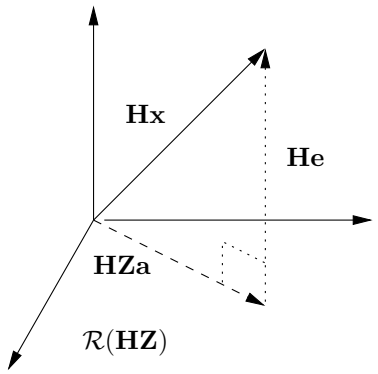

Fig. 1. Orthogonal projection of the filtered input onto the column space of the filtered signal model.

In the perceptual matching pursuit [23], which is a special case of the psychoacoustic adaptive matching pursuit with no adaptive norm, the dictionary element that minimizes the perceptual norm of the residual $\mathbf{r}_i$ is chosen. As in Section III, this is just the one-sinusoid nonlinear least-squares estimator operating on the residual. The matrix $\mathbf{Z}$ again reduces to the vector $\mathbf{z} = \left[\ \exp(j\omega 0)\ \cdots\ \exp(j\omega(N-1))\ \right]^T$, and the estimator is

$$\hat{\omega}_i = \underset{\omega}{\operatorname{argmin}} \|\mathbf{H}(\mathbf{r}_i - \mathbf{z}a)\|_2^2. \tag{43}$$

with $\mathbf{r}_i$ again being the residual at iteration i (see section III). Rewriting (43), we get the frequency estimator

$$\hat{\omega}_i = \underset{\omega}{\operatorname{argmax}}\, \mathbf{r}_i^H\mathbf{H}^2\mathbf{z}\left(\mathbf{z}^H\mathbf{H}^2\mathbf{z}\right)^{-1}\mathbf{z}^H\mathbf{H}^2\mathbf{r}_i \tag{44}$$

$$= \underset{\omega}{\operatorname{argmax}} \frac{|\ \langle \mathbf{H}\mathbf{z}, \mathbf{H}\mathbf{r}_i\rangle\ |^2}{\|\mathbf{H}\mathbf{z}\|_2^2}, \tag{45}$$

and the associated optimal scaling, i.e. amplitude and phase, is

$$\hat{a}_i = \frac{\langle \mathbf{H}\mathbf{z}, \mathbf{H}\mathbf{r}_i\rangle}{\|\mathbf{H}\mathbf{z}\|_2^2}. \tag{46}$$

The perceptual MP converges in the perceptual distortion measure rather than the 2-norm. We see that as with matching pursuit and the one-sinusoid NLS estimator, there is an equivalence between the perceptual matching pursuit and the perceptual NLS. The perceptual MP can be implemented efficiently using two FFTs in each iteration.

VI. EVD of the Perceptual Weighting Matrix

A. Signal Model Assumption

We now consider the example of a signal model component being an eigenvector $\mathbf{v}$ of the perceptual weighting matrix $\mathbf{H}$ with eigenvalue λ such that

$$\mathbf{H}\mathbf{v} = \lambda\mathbf{v}. \tag{47}$$

As we shall see in Section VII, this assumption leads to some interesting results and is indeed valid for certain important

cases. It is well-known that complex sinusoids are eigenvectors of convolution operators, i.e.

$$\mathbf{v} = [\ \exp(j\omega 0)\ \cdots\ \exp(j\omega(N-1))\ ]^T. \tag{48}$$

For notational simplicity, we omit the dependence of the eigenvalue λ on the frequency ω. Strictly speaking, (47) holds only in general (i.e. for any ω) for the asymptotic case $N \to \infty$. For the following analysis, consider (47) to be simply an approximation.

The above simplification requires the calculation of eigenvalues for the different eigenvector approximations. The optimal approximation of the eigenvalue for the vector $\mathbf{v}$ in a least-squares sense can be stated as

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}} \|\mathbf{H}\mathbf{v} - \lambda\mathbf{v}\|_2^2, \tag{49}$$

which is the Rayleigh coefficient, i.e.,

$$\hat{\lambda} = \frac{\mathbf{v}^H\mathbf{H}\mathbf{v}}{\mathbf{v}^H\mathbf{v}}. \tag{50}$$

We see that when the vector $\mathbf{v}$ is in fact an eigenvector of $\mathbf{H}$, this will result in the correct eigenvalue. The goodness of the eigenvalue approximation can conveniently be measured as

$$\|\mathbf{H}\mathbf{v} - \hat{\lambda}\mathbf{v}\|_2^2. \tag{51}$$

B. EVD of Circulant Matrices

In Section VI-A we considered the assumption that the signal model components are eigenvectors of the filtering matrix. Now we take a look at the eigenvalue decomposition of circulant matrices, i.e. the filtering matrix $\mathbf{H}$, which is also symmetric. A circulant matrix, say $\mathbf{C} \in \mathbb{R}^{M\times M}$, has the following structure

$$\mathbf{C} = \begin{bmatrix} c_0 & c_{M-1} & \cdots & c_1 \\ c_1 & c_0 & \cdots & c_{M-1} \\ \vdots & \vdots & \ddots & \vdots \\ c_{M-1} & c_{M-2} & \cdots & c_0 \end{bmatrix}, \tag{52}$$

which is uniquely defined by the vector $\mathbf{c} = [\ c_0\ \cdots\ c_{M-1}\ ]^T$. Defining the discrete Fourier transform (DFT) matrix as

$$\mathbf{F} = \frac{1}{\sqrt{M}}\left[\ \mathbf{f}_0 \quad \mathbf{f}_1 \quad \cdots \quad \mathbf{f}_{M-1}\ \right], \tag{53}$$

with the individual Fourier bases $\mathbf{f}_k = \left[\ f_k^0\ \cdots\ f_k^{M-1}\right]^T$ being composed from $f_k = \exp(j2\pi k/M)$. It then follows that the eigenvalue decomposition of the matrix $\mathbf{C}$ can be written as [36]

$$\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H, \tag{54}$$

with $\mathbf{U} = \mathbf{F}^H$ and $\mathbf{\Lambda} = \sqrt{M}\,\mathrm{diag}(\mathbf{F}\mathbf{c})$. We see that the eigenvalues in the diagonal matrix $\mathbf{\Lambda}$ are simply the DFT coefficients of $\mathbf{c}$ and the eigenvectors contained in $\mathbf{U}$ are the Fourier bases of a DFT. For the special case of a symmetric $\mathbf{c}$, i.e. $c_m = c_{M-m}$, the eigenvalues are real.

C. Equivalent Forms

We now use the EVD to write the perceptual distortion measure in a number of different but equivalent forms. First, we write the perceptual distortion as

$$D = \|\mathbf{H}(\mathbf{x} - \hat{\mathbf{x}})\|_2^2 = (\mathbf{x} - \hat{\mathbf{x}})^H\mathbf{H}^2(\mathbf{x} - \hat{\mathbf{x}}), \tag{55}$$

where $\mathbf{H}^2$ is also symmetric and circulant and has the eigenvalue decomposition $\mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^H$. Here it also interesting to note that comparing (55) to (15), we see that there is an inherent contradiction in the use of the perceptual weighting matrix and the inverse covariance matrix in the maximum likelihood estimator for the colored noise case since $\mathbf{H}^2 \neq \mathbf{\Sigma}^{-1}$. Now the perceptual weighting can be rewritten into the following diagonal form:

$$D = (\mathbf{x} - \hat{\mathbf{x}})^H\mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^H(\mathbf{x} - \hat{\mathbf{x}}) \tag{56}$$

$$= (\mathbf{U}^H\mathbf{x} - \mathbf{U}^H\hat{\mathbf{x}})^H\mathbf{\Lambda}^2(\mathbf{U}^H\mathbf{x} - \mathbf{U}^H\hat{\mathbf{x}}). \tag{57}$$

We note that the signal model $\hat{\mathbf{x}}$ may be chosen such that $\mathbf{U}^H\hat{\mathbf{x}}$ can be found analytically or pre-computed and stored in memory. Windowed sinusoids, for example, have simple Fourier transforms. As another example of this, we now treat the case of transform coding with the signal model components being equivalent to the eigenvectors, i.e. $\hat{\mathbf{x}} = \mathbf{U}\mathbf{y}$. In transform coding, the optimization problem concerns the transform coefficients $\mathbf{y}$. Bits are allocated such that the perceptual error is minimized. Now, the perceptual distortion can be rewritten as

$$D = (\mathbf{U}^H\mathbf{x} - \mathbf{y})^H\mathbf{\Lambda}^2(\mathbf{U}^H\mathbf{x} - \mathbf{y}), \tag{58}$$

or the equivalent form where the input signal $\mathbf{x}$ is pre-filtered:

$$D = \|\mathbf{H}\mathbf{x} - \mathbf{H}\mathbf{U}\mathbf{y}\|_2^2 = \|\mathbf{H}\mathbf{x} - \mathbf{U}\mathbf{\Lambda}\mathbf{y}\|_2^2. \tag{59}$$

It can be seen that distortion calculations can be simplified this way. This is a significant advantage in coding based on rate-distortion optimization [53], which requires the calculation of distortions for different allocations and quantizers.

VII. Relation to Simplified Estimators

A. Pre-filtering Method

Using the eigenvector assumption in (47) the sinusoidal frequency estimation criterion (38) can be significantly simplified:

$$\min \|\mathbf{H}(\mathbf{r}_i - \hat{\mathbf{r}}_i)\|_2^2 = \min \|\mathbf{H}\,(\mathbf{r}_i - \mathbf{v}a)\,\|_2^2 \tag{60}$$

$$= \min \|\mathbf{H}\mathbf{r}_i - \lambda\mathbf{v}a\|_2^2, \tag{61}$$

where a is a complex scale factor (amplitude and phase in polar form), which is included here since we do not restrict the norm or the phase of $\mathbf{v}$. The optimal value of this scale factor can then easily be found as

$$\hat{a} = \frac{\mathbf{v}^H\lambda^*\mathbf{H}\mathbf{r}_i}{\mathbf{v}^H|\lambda|^2\mathbf{v}}. \tag{62}$$

Next, expressing the perceptual NLS in terms of the unknown eigenvector, the frequency estimation criterion is simplified

significantly:

$$\hat{\omega}_i = \underset{\omega}{\operatorname{argmin}} \|\mathbf{H}\mathbf{r}_i - \lambda \mathbf{v} a\|_2^2 \tag{63}$$

$$= \underset{\omega}{\operatorname{argmax}} \frac{\mathbf{r}_i^H \mathbf{H}^H \lambda \mathbf{v}\mathbf{v}^H \lambda^* \mathbf{H}\mathbf{r}_i}{\mathbf{v}^H \lambda^* \lambda \mathbf{v}} \tag{64}$$

$$= \underset{\omega}{\operatorname{argmax}} \frac{|\langle \mathbf{v}, \mathbf{H}\mathbf{r}_i \rangle|^2}{N}. \tag{65}$$

We see that the estimator reduces to maximizing the inner product between the eigenvector $\mathbf{v}$ and $\mathbf{r}_i$ filtered by the perceptual filter. This inner product is just the periodogram of the perceptually filtered observed signal since $\mathbf{v}$ is a complex sinusoid. The modification of the signal model due to the filtering cancels out in the selection criterion and can be ignored. This is, however, not the case for damped sinusoids and pre-filtering is not well justified in that case. In practice this means that the input has to be filtered by the perceptual filter and then a squared error measure may be minimized in the estimation procedure if the model component is an eigenvector of $\mathbf{H}$ or is a reasonable approximation thereof.

The pre-filtering method has been applied to the perceptual estimation problem in e.g. [26], [54].

B. Pre- and Post-filtering Method

In the pre- and post-filtering approach of [55], [56], modeling is performed in the perceptual domain, i.e. operating on the pre-filtered signal:

$$\min \|\mathbf{H}\mathbf{r}_i - \hat{\mathbf{p}}\|_2^2 = \min \|\mathbf{p} - \hat{\mathbf{p}}\|_2^2. \tag{66}$$

Afterward, the modeled signal $\hat{\mathbf{p}}$ has to be mapped back to the signal domain by the inverse filter (also called the post-filter)

$$\hat{\mathbf{r}}_i = \mathbf{H}^{-1}\hat{\mathbf{p}}, \tag{67}$$

which means that the post-filter has to be sent to the decoder in coding applications. Otherwise, this approach differs from the pre-filtering method in algorithmic form in that the signal model is modified after the estimation/quantization rather than before. This has the advantage that the structure of the model, which may be lost by the filtering, is preserved in the estimation/quantization process. However, to argue that the signal model $\hat{\mathbf{p}}$ should be posed in the perceptual domain rather than in the signal domain seems somewhat contrived as the physical meaning of the model parameters is potentially lost in the transformation.

If the signal model component $\hat{\mathbf{p}}$ is an eigenvector of the inverse perceptual filter $\mathbf{H}^{-1}$, the post-filtering can be reduced to a simple scaling,

$$\hat{\mathbf{r}}_i = \lambda \hat{\mathbf{p}}, \tag{68}$$

in which case the signal model is valid also in the perceptual domain and can be modified directly. Also, the post-filter does not have to be transmitted to the receiver in this case.

For some types of estimators, though, the pre-filtering of the input signal has some serious drawbacks. Since it colors the signal, any noise will also be colored. The performance of subspace-based estimators degrades when the noise is not white [35]. Typically, this would be solved by applying pre-whitening but that is not an option for this application. These arguments favor NLS-based approximations such as matching pursuit for perceptual frequency estimation since NLS is still asymptotically efficient for colored noise [41].

C. Weighted Matching Pursuit

Since the filtering matrix $\mathbf{H}$ is symmetric, i.e. $\mathbf{H}^H = \mathbf{H}$, the inner product in the numerator of (65) can be written as

$$\langle \mathbf{v}, \mathbf{H}\mathbf{r}_i \rangle = \mathbf{v}^H \mathbf{H}\mathbf{r}_i = (\lambda \mathbf{v})^H \mathbf{r}_i, \tag{69}$$

such that the component selection criterion becomes

$$\hat{\omega}_i = \underset{\omega}{\operatorname{argmax}} \frac{|\langle \mathbf{v}, \mathbf{H}\mathbf{r}_i \rangle|^2}{N} = \underset{\omega}{\operatorname{argmax}} |\lambda|^2 \frac{|\langle \mathbf{v}, \mathbf{r}_i \rangle|^2}{N}. \tag{70}$$

The perceptual filtering approach can thus be reduced to a simple weighting of the inner products, where the weight is the absolute value of the eigenvalue associated with the eigenvector $\mathbf{v}$. This is in fact what the weighted matching pursuit does [25]. In the weighted MP the eigenvalue of a sinusoid of frequency ω is approximated as

$$\hat{\lambda} \approx \sqrt{A\left(\left\lfloor \frac{\omega K}{2\pi} + \frac{1}{2} \right\rfloor\right)}, \tag{71}$$

rather than the computationally more demanding least-squares approximation in (50). We see from (70) that under certain conditions on the perceptual filter, the sinusoidal estimator weighted MP is identical to the pre-filtering method. In [25], the weighting is introduced as a heuristic for incorporating psychoacoustics. Here, we have established the method as an approximation of the perceptual NLS.

The weighted MP has the problem that due to the perceptual weighting, the selected components may not be spectral maxima and spectral distortion introduced by the side-lobes of the sinusoidal components are not taken into account. This may cause audible artifacts. In the perceptual MP these problems are solved, and listening tests in [23] demonstrated its superior performance. The problems of the weighted MP can though easily be fixed by adding the constraints that the estimates have to be spectral maxima.

VIII. Numerical Examples

In this section we illustrate some of the points made in the previous sections using an example of a sinusoidal audio signal, the trumpet signal of SQAM [57]. In Figure 2 a segment of this signal is shown. The signal is sampled at 44.1 kHz. The masking curve is derived using the model in [30] and the corresponding perceptual weighting function is shown in Figure 3 along with the periodogram of the segment in Figure 2. Note the very distinct peaks and the harmonic structure in the periodogram.

The convergence of the perceptual MP in the perceptual norm is illustrated in Figure 4, again for the trumpet signal in Figure 2. Note how the perceptual distortion is a non-increasing function of the number of components. The sinusoidal frequencies that are estimated in the individual iterations of the perceptual MP (indicated by numbers) are shown in Figure 5. The effect of the perceptual distortion measure can

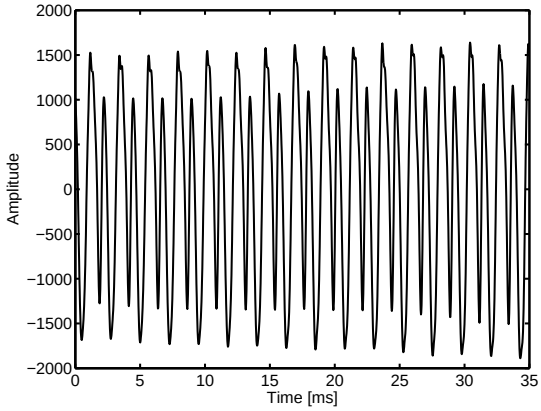

Fig. 2. Example of an audio segment, trumpet. The trumpet signal is a fairly stationary, tonal signal.

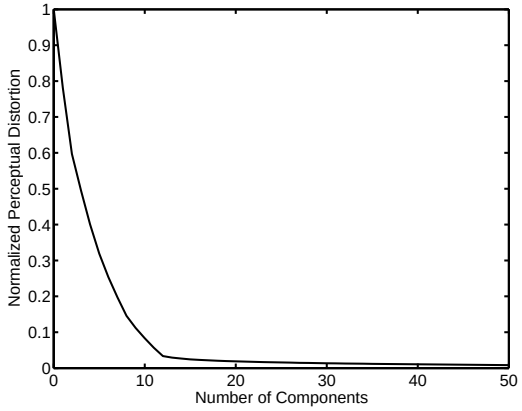

Fig. 4. Convergence of the perceptual matching pursuit in the perceptual distortion for the trumpet signal in Figure 2.

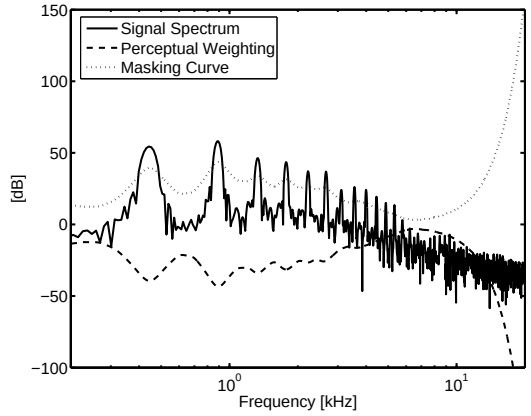

Fig. 3. Perceptual weighting function (dashed), masking curve (dotted) and spectrum for the trumpet signal (solid) in Figure 2.

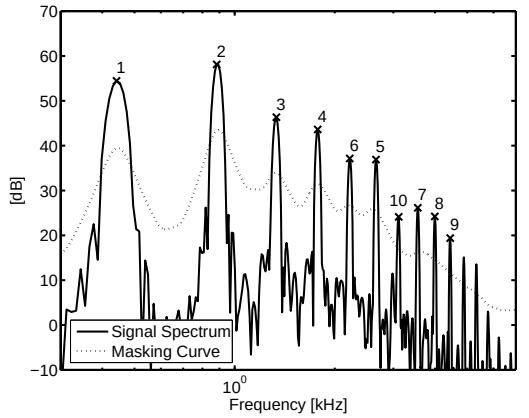

Fig. 5. Frequencies estimated (crosses) in the individual iterations (indicated by number) by the perceptual matching pursuit.

been observed in that although more energy is present at peak 2, the perceptual MP picks peak 1 first. From the figure it is clear that the effect of the perceptual distortion measure is one of ordering. In Figure 6 an illustration of the error introduced by the eigenvector/-value approximation is shown. The figure shows the perceptual weighting for a segment of the trumpet signal and the error as defined by (51) introduced as a function of frequency with the eigenvalues being approximated using (50). Also shown is the signal-to-noise ratio (SNR), which is calculated as

$$SNR = 10 \log_{10} \frac{\|\mathbf{Hv}\|_2^2}{\|\mathbf{Hv} - \hat{\lambda}\mathbf{v}\|_2^2} [\text{dB}]. \tag{72}$$

The perceptual weighting was derived with a frequency resolution of 4096 uniformly spaced points, and the corresponding filter was calculated by taking the inverse discrete Fourier transform of its square-root. The complex sinusoids were windowed by a Hanning window having a length of 1544 samples and then zero-padded to length 4096 to match the length of the perceptual filter. These are fairly typical choices of constants in audio coding. From this figure, it is very clear that these windowed, zero-padded complex sinusoids are not eigenvectors of the filtering matrix, since the SNR is far from the numerical noise floor (64 bit floating point). The loss in estimator performance in terms of distortion may well be worth it, though, as considerable complexity reductions can be achieved. It can also be seen that the goodness of the approximation is highly frequency dependent with the approximation performing well at high frequencies for this particular perceptual weighting function. This can be attributed to the perceptual weighting function being flatter in this region. Note that the perceptual weighting function will be dominated by the threshold in quiet for very low and high frequencies. When the length of the perceptual filter and the complex sinusoids are the same and no window is applied, the error hits the numerical noise floor as the complex sinusoids become eigenvectors of the filtering matrix.

IX. Results

In this section we briefly summarize and discuss the main results of this paper. In particular we recapitulate the conditions under which the different methods that have been discussed are equivalent and optimal.

- When estimating the frequencies of sinusoids in additive white and Gaussian noise, the nonlinear least-squares method is the maximum likelihood estimator. The nonlinear least-squares method is efficient, i.e. in this case it attains the Cramér-Rao bound and is hence optimal.

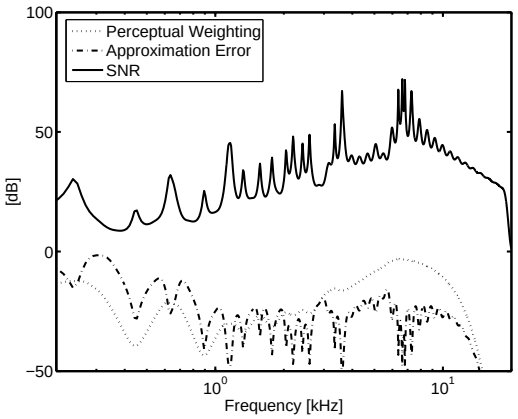

Fig. 6. The error (solid) and SNR (dashed) as function of frequency introduced by the eigenvector approximation for a particular perceptual weighting function (dotted).

- Under the condition that the noise is Gaussian but colored, there is an equivalence between maximum likelihood and (weighted) least-squares based estimation. The non-linear least-squares method is still asymptotically optimal in this case.
- The matching pursuit algorithm is a (one-dimensional) relaxation of the subspace pursuit of nonlinear least-squares. It converges in the respective norm, here the 2-norm, as a function of the number of components and has an efficient implementation for the frequency estimation problem.
- A recently established perceptual distortion measure, which shapes the error spectrum according to the masking threshold, can be shown to form a circulant and symmetric perceptual weighting matrix, which can be interpreted as a filtering matrix. Circulant and symmetric weighting matrices have eigenvectors that are rectangularly windowed complex sinusoids of uniformly spaced frequencies. Asymptotically, sinusoids of arbitrary frequencies are eigenvectors of the weighting matrix.
- When this perceptual weighting matrix is applied in solving the least-squares problems in the NLS and MP estimators, we get the perceptual nonlinear least-squares estimator and the simpler perceptual matching pursuit. The perceptual matching pursuit now converges in the perceptual distortion.
- The pre-filtering method and the weighted matching pursuit are equivalent to the perceptual matching pursuit when the model components are eigenvectors of the perceptual weighting matrix. This allows for very efficient implementation of the perceptual weighting. In some applications of the pre-filtering method and the weighted matching pursuit, the model components are not eigenvectors of the weighting matrix; then, these methods are only approximations of the perceptual matching pursuit.

X. Conclusion

We have introduced the perceptual frequency estimation problem based on a spectral distortion measure and its optimal solution, the nonlinear least-squares frequency estimator. The nonlinear least-squares method has a strong background in statistical signal processing and estimation theory and is well-known to have excellent statistical performance. We have related this to a number of well-known methods for perceptual parameter estimation, namely the perceptual matching pursuit, the weighted matching pursuit and the pre-filtering method. It has been shown that these methods can be seen as relaxations and approximations of the optimal solution. Specifically, we have established the perceptual matching pursuit as a relaxation of the nonlinear least-squares estimator, and we have shown that it reduces to the pre-filtering method and the weighted matching pursuit under certain conditions.

XI. Acknowledgment

The authors would like to thank Christoffer A. Rødbro, Dept. of Communication Technology, Aalborg University, Denmark as well as the anonymous reviewers for their constructive comments that helped improve this manuscript.

References

[1] P. Hedelin, "A tone oriented voice excited vocoder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 6, Apr. 1981, pp. 205–208.

[2] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B.V., 1995, ch. 4.

[3] ——, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34(4), pp. 744–754, Aug. 1986.

[4] ——, "Speech Transformation Based on a Sinusoidal Representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34(6), pp. 1449–1464, Dec. 1986.

[5] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. Speech and Audio Processing*, vol. 5(5), pp. 389–406, Sept. 1997.

[6] J. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. Speech and Audio Processing*, vol. 9(7), pp. 731–740, Oct. 2001.

[7] J. Lindblom, "A sinusoidal voice over packet coder tailored for the frame-erasure channel," *IEEE Trans. Speech and Audio Processing*, 2004, accepted.

[8] C. A. Rødbro, M. G. Christensen, S. H. Jensen, and S. V. Andersen, "Compressed domain packet loss concealment of sinusoidally coded speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, Apr. 2003, pp. 104–107.

[9] C. A. Rødbro, J. Jensen, and R. Heusdens, "Rate-distortion optimal time-segmentation and redundancy selection for VoIP," *IEEE Trans. Speech and Audio Processing*, 2005, accepted.

[10] C. A. Rødbro, M. N. Murthi, S. V. Andersen, and S. H. Jensen, "Hidden Markov Model Based Framework for Packet Loss Concealment in Voice over IP," *IEEE Trans. Speech and Audio Processing*, 2005, accepted.

[11] E. B. George and M. J. T. Smith, "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis-synthesis of musical tones," *J. Audio Eng. Soc.*, vol. 40(6), pp. 497–516, June 1992.

[12] J. O. Smith III and X. Serra, "PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation," in *Proc. International Computer Music Conference*, Aug. 1987, pp. 290–297.

[13] J. O. Smith III and X. Serra, "Spectral Modelling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic Plus Stochastic Decomposition," *Computer Music Journal*, vol. 14(4), pp. 12–24, Winter 1990.

[14] K. N. Hamdy, M. Ali, and A. H. Tewfik, "Low bit rate high quality audio coding with combined harmonic and wavelet representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 1996, pp. 1045–1048.

[15] R. Heusdens and S. van de Par, "Rate-distortion optimal sinusoidal modeling of audio using psychoacoustical matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2002, pp. 1809–1812.

[16] B. Edler, H. Purnhagen, and C. Ferekidis, "ASAC – Analysis/Synthesis Audio Codec for Very Low Bit Rates," in *100th Conv. Aud. Eng. Soc.*, May 1996, paper preprint 4179.

[17] ISO/IEC, *Coding of audio-visual objects – Part 3: Audio (MPEG-4 Audio Edition 2001)*. ISO/IEC Int. Std. 14496-3:2001, 2001.

[18] E. G. P. Schuijers, A. W. J. Oomen, A. C. den Brinker, and J. Breebart, "Advances in parametric coding for high-quality audio," in *114th Conv. Aud. Eng. Soc.*, Feb. 2003, paper preprint 5852.

[19] T. S. Verma and T. H. Y. Meng, "A 6kbps to 85kbps scalable audio coder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, June 2000, pp. 877–880.

[20] S. N. Levine and J. O. Smith III, "A switched parametric & transform audio coder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, Mar. 1999, pp. 985–988.

[21] K. Vos, R. Vafin, R. Heusdens, and W. B. Kleijn, "High-quality consistent analysis-synthesis in sinusoidal coding," in *Proc. Audio Eng. Soc. 17th Conf: High Quality Audio Coding*, Sept. 1999, pp. 244–250.

[22] R. Vafin, S. V. Andersen, and W. B. Kleijn, "Exploiting time and frequency masking in consistent sinusoidal analysis-synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, June 2000, pp. 901–904.

[23] R. Heusdens, R. Vafin, and W. B. Kleijn, "Sinusoidal modeling using psychoacoustic-adaptive matching pursuits," *IEEE Signal Processing Lett.*, vol. 9(8), pp. 262–265, Aug. 2002.

[24] K. Hermus, W. Verhelst, P. Lemmerling, P. Wambacq, and S. van Huffel, "Perceptual audio modeling with exponentially damped sinusoids," *Signal Processing*, vol. 85(1), pp. 163–176, Jan. 2005.

[25] T. S. Verma and T. H. Y. Meng, "Sinusoidal modeling using frame-based perceptually weighted matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, Mar. 1999, pp. 981–984.

[26] J. Jensen, R. Heusdens, and S. H. Jensen, "A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids," *IEEE Trans. Speech and Audio Processing*, vol. 12(2), pp. 121–132, Mar. 2004.

[27] H. Purnhagen, N. Meine, and B. Edler, "Sinusoidal coding using loudness-based component selection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, May 2002, pp. 1817–1820.

[28] T. Painter and A. S. Spanias, "Perceptual segmentation and component selection for sinusoidal representations of audio," *IEEE Trans. Speech and Audio Processing*, vol. 13(2), pp. 149–162, Mar. 2005.

[29] B. Atal and M. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27(3), pp. 247–254, June 1979.

[30] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens, "A new psychoacoustical masking model for audio coding applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, May 2002, pp. 1805 – 1808.

[31] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. on Applied Signal Processing*, vol. 9, pp. 1292–1304, 2005.

[32] R. Heusdens, J. Jensen, W. B. Kleijn, V. kot, O. A. Niamut, S. van de Par, N. H. van Schijndel, and R. Vafin, "Bit-rate scalable intra-frame sinusoidal audio coding based on rate-distortion optimisation," *J. Audio Eng. Soc.*, 2005, submitted.

[33] M. G. Christensen and S. van de Par, "Efficient parametric coding of transients," *IEEE Trans. Speech and Audio Processing*, 2005, accepted.

[34] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, 1993.

[35] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Prentice Hall, 1997.

[36] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. The Johns Hopkins University Press, 1996.

[37] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: Survey, new results and an application," *IEEE Trans. Signal Processing*, vol. 48(2), pp. 338–352, Feb. 2000.

[38] M. M. Goodwin, "Matching pursuit and atomic signal models based on recursive filter banks," *IEEE Trans. Signal Processing*, vol. 47(7), pp. 1890–1902, July 1999.

[39] G. Davis, "Adaptive nonlinear approximations," Ph.D. dissertation, New York University, Sept. 1994.

[40] P. Stoica, R. L. Moses, B. Friedlander, and T. Söderström, "Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37(3), pp. 378–392, Mar. 1989.

[41] P. Stoica, A. Jakobsson, and J. Li, "Cisiod Parameter Estimation in the Coloured Noise Case: Asymptotic Cramér-Rao Bound, Maximum Likelihood, and Nonlinear Least-Squares," in *IEEE Trans. Signal Processing*, vol. 45(8), Aug. 1997, pp. 2048–2059.

[42] J. Li and P. Stoica, "Efficient mixed-spectrum estimation with application to target feature extraction," *IEEE Trans. Signal Processing*, vol. 44(2), pp. 281–295, Feb. 1996.

[43] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41(12), pp. 3397–3415, Dec. 1993.

[44] M. M. Goodwin, "Adaptive Signal Models: Theory, Algorithms, and Audio Applications," Ph.D. dissertation, University of California, Berkeley, 1997.

[45] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with application to wavelet decomposition," in *Conference Record of the Twenty-Seventh Asilomar Conference on Signals, Systems, and Computers*, vol. 1, Nov. 1993, pp. 40–44.

[46] H. Feichtinger, A. Turk, and T. Strohmer, "Hierarchical parallel matching pursuit," *Proc. SPIE: Image Reconstruction and Restoration*, vol. 2302, pp. 222–232, July 1994.

[47] J. Adler, B. Rao, and K. Kreutz-Delgado, "Comparison of basis selection methods," in *Conference Record of the Thirtieth Asilomar Conference on Signals, Systems, and Computers*, vol. 1, Nov. 1996, pp. 252–257.

[48] E. B. George and M. J. T. Smith, "A new speech coding model based on a least-squares sinusoidal representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 12, Apr. 1987, pp. 1641–1644.

[49] J. Plasberg, D. Zhao, and W. B. Kleijn, "Sensitivity matrix for a spectro-temporal auditory model," in *Proc. XII European Signal Processing Conf. (EUSIPCO)*, Sept. 2004, pp. 1673–1676.

[50] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Am.*, vol. 99(6), pp. 3615–3622, June 1996.

[51] ——, "A quantitative model of the effective signal processing in the auditory system. II. Simulations and measurements," *J. Acoust. Soc. Am.*, vol. 99(6), pp. 3623–3631, June 1996.

[52] T. Painter and A. S. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88(4), pp. 451–515, Apr. 2000.

[53] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36(9), pp. 1445–1453, Sept. 1988.

[54] R. Hendriks, R. Heusdens, and J. Jensen, "Perceptual linear predictive noise modelling for sinusoid-plus-noise audio coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, May 2004, pp. 189–192.

[55] B. Edler and G. Schuller, "Audio coding using a psychoacoustic pre- and post-filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, June 2000, pp. 881–884.

[56] G. D. T. Schuller, B. Yu, D. Huang, and B. Edler, "Perceptual audio coding using adaptive pre- and post-filters and lossless compression," in *IEEE Trans. Speech and Audio Processing*, vol. 10(6), Sept. 2002, pp. 379–390.

[57] European Broadcasting Union, *Sound Quality Assessment Material Recordings for Subjective Tests*. EBU, Apr. 1988, Tech. 3253.